

## METHODOLOGY ARTICLE

## Open Access

# Genetic distance as an alternative to physical distance for definition of gene units in association studies

Cristina Rodriguez-Fontenla, Manuel Calaza and Antonio Gonzalez\*

## Abstract

**Background:** Some association studies, as the implemented in VEGAS, ALIGATOR, i-GSEA4GWAS, GSA-SNP and other software tools, use genes as the unit of analysis. These genes include the coding sequence plus flanking sequences. Polymorphisms in the flanking sequences are of interest because they involve cis-regulatory elements or they inform on untyped genetic variants through linkage disequilibrium. Gene extensions have customarily been defined as  $\pm 50$  Kb. This approach is not fully satisfactory because genetic relationships between neighbouring sequences are a function of genetic distances, which are only poorly replaced by physical distances.

**Results:** Standardized recombination rates (SRR) from the deCODE recombination map were used as units of genetic distances. We searched for a SRR producing flanking sequences near the  $\pm 50$  Kb offset that has been common in previous studies. A  $SRR \geq 2$  was selected because it led to gene extensions with median length = 45.3 Kb and the simplicity of an integer value. As expected, boundaries of the genes defined with the  $\pm 50$  Kb and with the  $SRR \geq 2$  rules were rarely concordant. The impact of these differences was illustrated with the interpretation of top association signals from two large studies including many hits and their detailed analysis based in different criteria. The definition based in genetic distance was more concordant with the results of these studies than the based in physical distance. In the analysis of 18 top disease associated loci from the first study, the  $SRR \geq 2$  genes led to a fully concordant interpretation in 17 loci; the  $\pm 50$  Kb genes only in 6. Interpretation of the 43 putative functional genes of the second study based in the  $SRR \geq 2$  definition only missed 4 of the genes, whereas the based in the  $\pm 50$  Kb definition missed 10 genes.

**Conclusions:** A gene definition based on genetic distance led to results more concordant with expert detailed analyses than the commonly used based in physical distance. The genome coordinates for each gene are provided to maintain a simple use of the new definitions.

## Background

Genes are the unit of analysis or interpretation of multiple genetic association studies. However, multiple operational definitions of genes coexist in current use. Some are restricted to the coding sequence but, most often, they are extended to include flanking sequences because they contain polymorphisms that are informative of variation in the coding sequence through linkage disequilibrium (LD) or polymorphisms that are themselves functional by involving regulatory sequences. Here, we

have addressed the definition of these gene extensions for application in gene- or pathway-based association studies, gene-based interaction analysis and interpretation of large numbers of top association signals for meta-analysis or for gene- and pathway- enrichment analysis.

Gene- or pathway- based association studies [1-8] consider the genes, not the individual SNPs, as the units of analysis. Association statistics for the genes are obtained by combining the statistics corresponding to the SNPs mapping to each of them. In this way, it becomes possible to identify genes with multiple independent SNPs contributing to the trait but lacking significant association on their own. The same considerations apply to pathway- or gene-set analyses, where the association

\* Correspondence: [antonio.gonzalez.martinez-pedrayo@sergas.es](mailto:antonio.gonzalez.martinez-pedrayo@sergas.es)  
Laboratorio de Investigacion 10 and Rheumatology Unit, Instituto de Investigacion Sanitaria - Hospital Clinico Universitario de Santiago, Santiago de Compostela, Spain

signals from the genes in a pathway are combined. A similar situation appears in interaction analyses where the objective is to identify pairs of genes contributing to a trait in a way that deviates from the simple addition of their independent effects [9,10]. This type of analysis can be done at the individual SNP level but this is very sensitive to small variations in the study, and analysis at the gene or pathway level has been advocated as more reproducible [9-11]. In addition, extended gene definitions can be useful in analysis that by considering many top association signals find it impractical a detailed analysis of each of them. For example, when it is necessary to decide if associations from a large number of studies are coincident or not in the same gene [12], or when interpreting multiple association signals [13,14].

In all these situations, genes have been operationally defined as the coding sequence plus a fixed physical distance in each direction. Length of the extensions has been from 0 to 500 [5,7] Kb, but most often of 20 [8,13,14] or 50 [1-4,9] Kb. This is a practical solution that is used because of its simplicity, but this definition is subjective and not fit for many genes. Here, we propose a definition of genes that is equally easy to apply and has the advantage of including genetic distance in place of physical distance. Genetic distance is the relevant one because it determines LD between polymorphisms [15-18] and, therefore, the information that SNPs in the extensions provide about un-typed variation in the coding or regulatory sequences. Genetic and physical distances are not interchangeable because the correspondence between the two is very variable along the genome [15-18]. We took genetic distances as standardized

recombination ratios (SRR) from the deCODE recombination map [16], which is the most accurate available. The new extended gene definitions were compared with definitions based on physical distances to illustrate their advantages. They are made available in a text file with genome coordinates to facilitate their use.

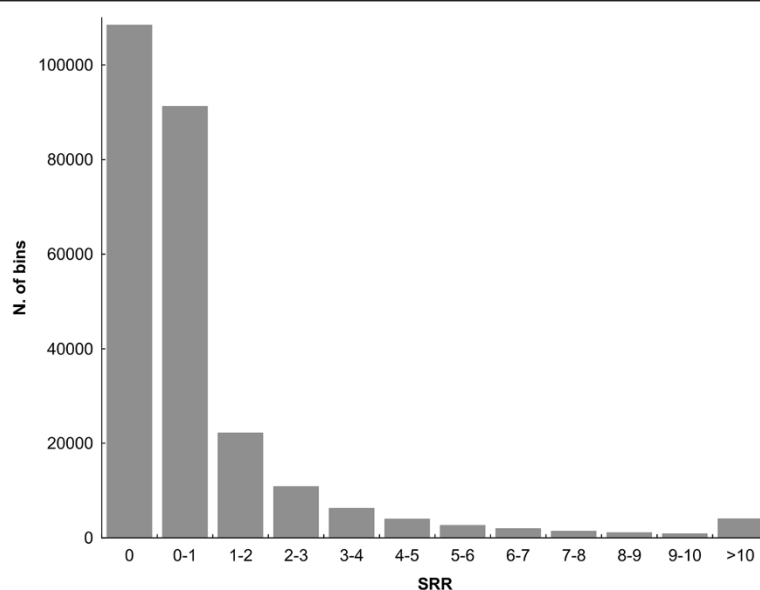
## Results

### Setting a SRR threshold

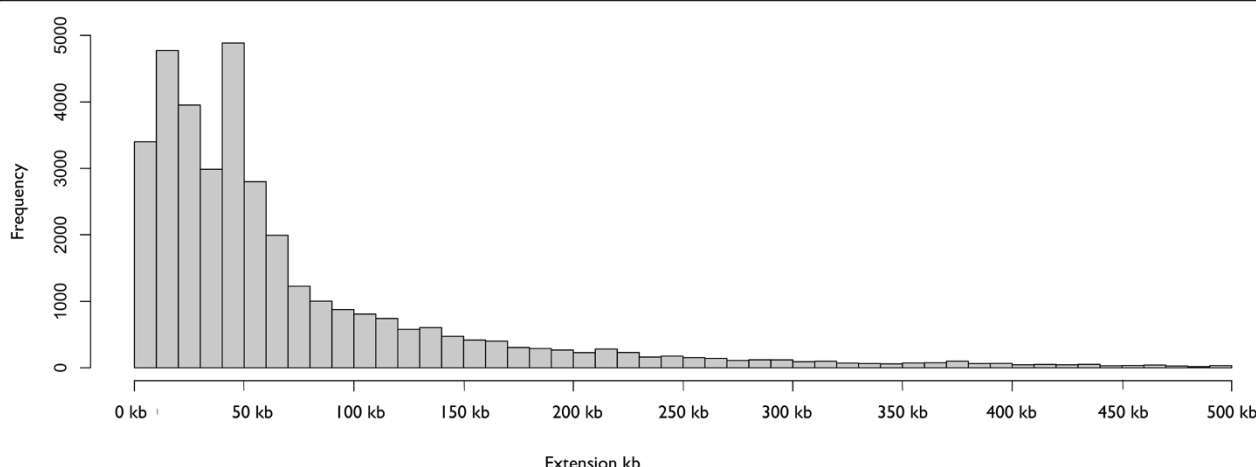
It is well known that the recombination rate is very irregular along the human genome [15-18]. This irregularity leads to a skewed distribution of SRR along the genome (Figure 1) [16] including a large fraction of bins, 42.6%, with no recombination ( $SRR = 0$ ) and 78.4% of the bins with less than the average ( $SRR < 1$ ). Therefore, most of the recombination takes place in the remaining 21.6% bins. Analysis of the SRR distribution showed that extensions of genes based on an  $SRR \geq 2$  have a median physical length of 45.3 Kb (IQR = 22.9-90.2 Kb). This median length is similar to the most common physical distance extension used until now, which is of 50 Kb. The  $SRR \geq 2$  is only found in a minor fraction of bins, 12.9%. The remaining 87.1% of the 10 Kb bins showed lower SRR. No detailed optimization of the SRR was attempted preferring to keep the simplicity of an integer value.

### Comparison of genetic and physical distance based gene definitions

Concordance between the median length of the extensions based on  $SRR \geq 2$  and the  $\pm 50$  Kb rule made possible a direct comparison. However, the new definitions obtained here account for recombination and are variable (Figure 2),



**Figure 1** Distribution of the standardized recombination rate (SRR) in the human genome. Number of 10 Kb bins from the deCODE recombination map [16] within each interval of SRR values.



**Figure 2 Length distribution of the 36 044 gene extensions according to the  $SRR \geq 2$  rule.** The 5' and 3' extensions for each gene have been separately considered. All followed the  $SRR \geq 2$  rule except for 2669 of genes near telomeres and centromeres, where information is incomplete and that were replaced by the median length of extensions in their chromosomes; most of them in the 40–50 Kb range.

not uniform. They go from less than 10 Kb (8.8% of the extensions) to more than 500 Kb (1.2% of the extensions). The distribution of extension lengths implies that most gene boundaries are discordant between the two definitions. In fact, only 21.3% of the extensions obtained with one definition are within  $\pm 10$  Kb of the obtained with the other, and even less frequently (6.1%) when the two extensions of a gene are considered simultaneously.

We have used two large GWAS with multiple associated loci to illustrate differences between the two gene definitions. However, these analyses should not be confused with an attempt to replace detailed analysis of GWAS results. First, we used the interpretation of 18 top association signals from the 2007 WTCCC GWAS [19]. The authors of this study gave lists of relevant genes for each associated locus based on analysis of the associated SNPs and LD around the top signal. These lists include from 0 to 23 genes. The  $SRR \geq 2$  definition led to lists that were more concordant with the WTCCC GWAS than the obtained with the  $\pm 50$  Kb definition (Table 1 and Additional file 1: Table S1 [20]). All the genes selected by the WTCCC authors were also included when applying the two definitions, but in some loci the gene definitions led to consider some extra genes. Specifically, the  $SRR \geq 2$  definition included additional genes in one locus, whereas the  $\pm 50$  Kb definition included additional genes in 12 of the 18 loci ( $P = 0.00015$  for the comparison of fully concordant loci). In more detail, six loci included an extra gene according to the  $\pm 50$  Kb rule (an example shown in Figure 3A); four loci included two extra genes with the  $\pm 50$  Kb definition (two of these loci shown in Figure 3B and C); an additional locus included 3 extra genes in the list obtained with the  $\pm 50$  Kb definition (Table 1). The remaining locus was the

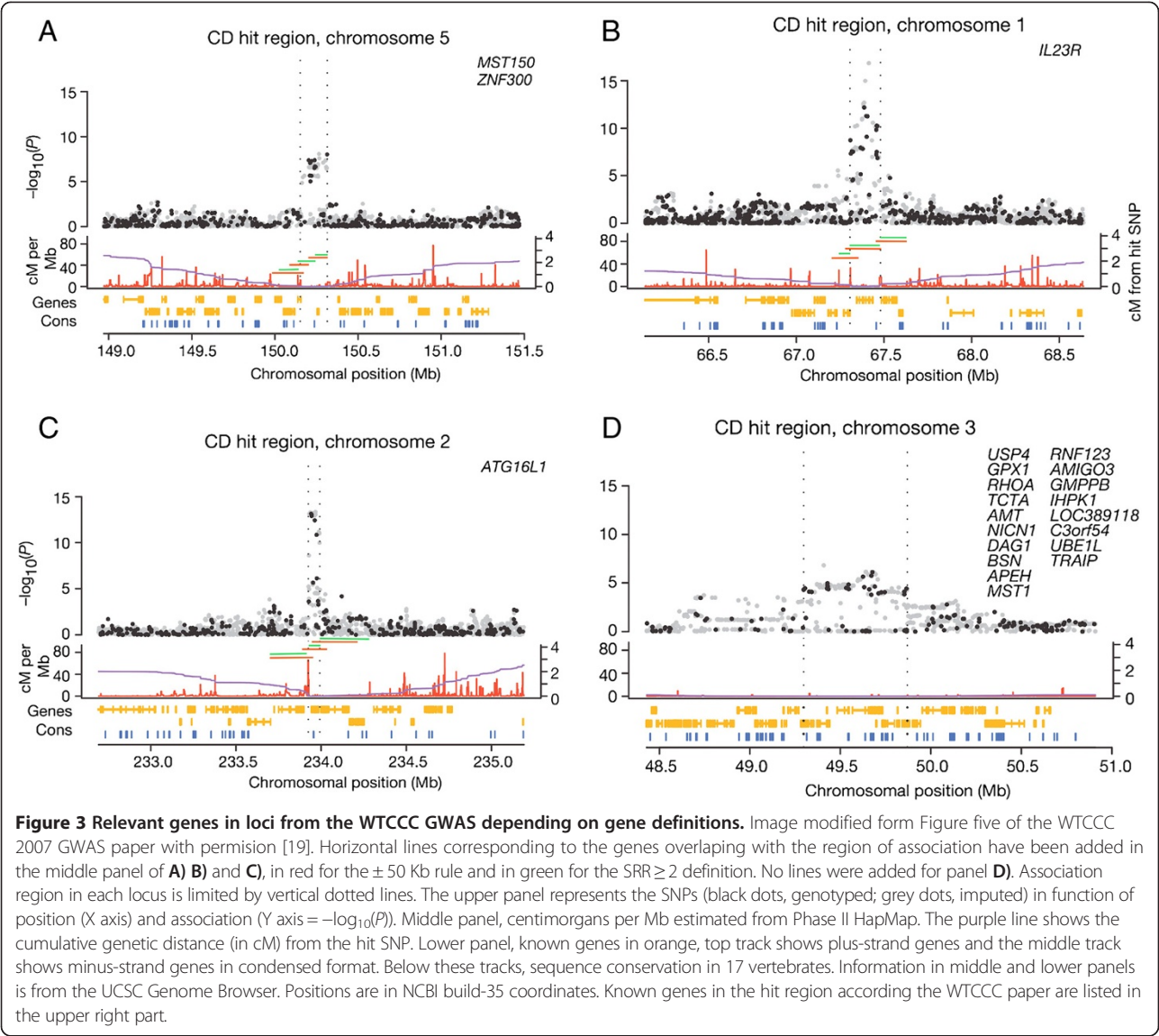
**Table 1 Number of genes in association regions of the WTCCC GWAS top hits [19]**

Chromosome	Disease <sup>b</sup>	WTCCC	$SRR \geq 2^c$	$\pm 50$ Kb <sup>c</sup>
5p13	CD	0	-	-
10q24	CD	1	-	-
10q25	T2D	1	-	-
9p21	CAD	2	-	-
10q21	CD	3	-	-
16q12	CD	4	-	-
16q12	T2D	1	-	+ 1
5q33	CD	2	-	+ 1
1p13 <sup>a</sup>	RA	7	-	+ 1
1p13 <sup>a</sup>	T1D	7	-	+ 1
16p13	T1D	8	-	+ 1
16p12	BD	9	-	+ 1
1p31	CD	1	-	+ 2
2q37	CD	1	-	+ 2
18p11	CD	1	-	+ 2
12q24	T1D	15	-	+ 2
12q13	T1D	26	-	+ 3
3p21	CD	18	+ 7	+ 9
Total		107	+ 7	+ 26

<sup>a</sup>These two loci overlap.

<sup>b</sup>CD = Crohn's disease, T2D = Type 2 diabetes, CAD = Coronary artery disease, RA = Rheumatoid arthritis, T1D = Type 1 diabetes, BD = Bipolar disorder.

<sup>c</sup> $SRR \geq 2$  for the gene definition extended to reach a cumulative  $SRR \geq 2$  in each direction; and  $\pm 50$  Kb for gene definition extended to this length in each direction. Only changes in the number of genes, not in their identity, were observed between the three lists: no differences with the genes highlighted by the WTCCC authors; + number of additional genes beyond the highlighted by the WTCCC authors. A full list of genes in each loci is available as Additional file 1: Table S1.



unique in which the three lists were discordant. This locus is particularly difficult because it shows a very low recombination rate and, therefore, a very wide region of association with ill-defined limits (Figure 3D). In addition, it shows a high density of genes implying large differences when applying alternative criteria. Overall, there were 107 genes in the 18 association regions according with detailed analysis done by the WTCCC authors. The definition based on genetic distances led to fully concordant results except for the difficult locus, where no criterion can be considered certain (Figure 3D). In contrast, the definition based on  $\pm 50$  Kb included 26 additional genes ( $P = 0.00025$  for the comparison of the number of extra genes). Nine of these extra genes were from the difficult locus in chromosome 3, but there were 17 extra genes in other loci. This example illustrates the very

good concordance between *post-hoc* detailed analysis of each locus done by the WTCCC authors and the simple overlap with gene definitions based on genetic distances. It also illustrates the differences between this definition and the based on a fixed physical distance.

The second study used to illustrate differences between the gene definitions is a large GWAS that included a selection of putative functional candidate genes for many of the associated loci [21]. The authors of this study used two criteria to identify these genes. The two were based in SNPs that are in high LD ( $r^2 > 0.8$ ) with the top associated SNP and with predictable functional relevance because they disrupt the protein sequence, nsSNPs, or the expression of a nearby gene, *cis*-eQTLs. The search extended to the more than 3 000 genes mapping 1 Mb around the 75 top associated signals. It led to 43 functional candidates

**Table 2 Functional candidate genes that are missed depending on the gene definition**

Chromosome <sup>a</sup>	Phenotype <sup>b</sup>	van der Harst <i>et al.</i>	SRR $\geq 2^c$	$\pm 50$ Kb <sup>c</sup>
nsSNP				
1q23	MCHC	<i>OR6Y1, OR10Z1, SPTA1</i>	-	<i>OR6Y1</i>
1q44	RBC	<i>TRIM58</i>	-	-
6p21	MCH	<i>HFE</i>	<i>HFE</i>	<i>HFE</i>
6p21	RBC	<i>HLA-DQA1</i>	-	-
10q11	MCV	<i>MARCH8</i>	-	-
11q13	MCV	<i>RPS6KB2</i>	-	-
11q13	HB	<i>ARHGEF17</i>	-	-
12q24	HB	<i>SH2B3</i>	-	-
12q24	MCV	<i>ACADS</i>	-	-
16q22	RBC	<i>CTRL, PSMB10</i>	-	<i>CTRL, PSMB10</i>
19p13	MCV	<i>UBXD1, NUDT19</i>	<i>UBXD1</i>	<i>UBXD1</i>
22q11	MCV	<i>YDJC</i>	<i>YDJC</i>	-
22q12	MCH	<i>FBXO7, TMPPRS6</i>	-	-
	$\Sigma$ nsSNP	18	-3	-5
eQTL				
4q27	MCV	<i>CCNA2</i>	-	-
6p23	MCH	<i>GMPR</i>	-	-
6p21	RBC	<i>HLA-DQA1/HLA-DQA2</i>	<i>HLA-DQA2</i>	<i>HLA-DQA2</i>
8p11	MCHC	<i>C8orf40</i>	-	-
10q11	MCV	<i>MARCH8</i>	-	-
11p15	HB	<i>AKIP1/C11orf16, NRIP3</i>	-	<i>NRIP3</i>
11q13	MCV	<i>RPS6KB2, PTPRCAP/COROB1</i>	-	-
11q13	HB	<i>ARHGEF17</i>	-	-
15q22	MCV	<i>PTPLAD1</i>	-	<i>PTPLAD1</i>
15q25	MCHC	<i>DNAJA4</i>	-	-
16q22	RBC	<i>DUS2L</i>	-	<i>DUS2L</i>
17q11	MCH	<i>ERAL1, TRAF4</i>	-	<i>ERAL1</i>
17q12	RBC	<i>CDK12</i>	-	-
17q25	HB	<i>PGS1</i>	-	-
18q21	MCH	<i>C18orf25</i>	-	-
19p13	MCH	<i>CALR, FARSA</i>	-	-
22q11	MCV	<i>UBE2L3</i>	-	-
22q13	MCV	<i>ECGF1</i>	-	-
	$\Sigma$ eQTL	25	-1	-5
	Total	43	-4	-10

<sup>a</sup>Loci in chromosome 17q21 were excluded from analysis because it contains a common inversion polymorphism of approximately 900 kb in populations with European ancestry that shows exceptional LD and inheritance [26].

<sup>b</sup>Phenotypes were: *MCHC* = Mean cell haemoglobin concentration, *RBC* = Red blood cell count, *MCH* = Mean cell haemoglobin, *MCV* = Mean cell volume and *HB* = Haemoglobin.

<sup>c</sup>Genes that did not overlap with the SRR  $\geq 2$  or the  $\pm 50$  Kb definition are indicated: - no differences with the functional candidate genes highlighted by van der Harst *et al.* [20]; genes that were highlighted by van der Harst *et al.* [20] but whose definition did not overlap with the top associated SNP.

Functional candidates were selected in van der Harst *et al.* [20] because they contained nsSNP (upper rows) or were regulated by eQTL (lower rows) in LD with the top associated SNP.



(Table 2). These putative functional candidates were prioritized relative to other genes in the loci and the aim of our current test has been to evaluate the capacity of the two gene definitions to highlight them. We found that the  $SRR \geq 2$  definition performed better than the definition based on  $\pm 50$  Kb (Table 2). The difference was due to a larger number of genes failing to be highlighted by the latter approach. In more detail: the two methods missed the same candidate genes in 3 loci, the  $SRR \geq 2$  definition missed an additional candidate, but the  $\pm 50$  Kb missed other 7 candidate genes ( $P = 0.028$ ; Table 2). In this way, the  $SRR \geq 2$  definition missed 9.3% of the putative functional candidates, whereas the  $\pm 50$  Kb definition missed 23.3% of them.

## Discussion

The gene definitions based on genetic distances lead to extensions with different physical lengths, meaning that most gene definitions are discordant from any other based on a fixed length as we have shown for the  $SRR \geq 2$  and  $\pm 50$  Kb. The advantages of the new definition stem from the fact that physical distance is an inaccurate substitute of genetic distance as a measure of the relationships between polymorphisms in the population [15-18]. This has been illustrated by showing a better performance in the interpretation of top association signals of the simple overlap rule based on  $SRR \geq 2$  definitions than in the traditional  $\pm 50$  Kb. In consequence, the new gene definitions will improve gene- and pathway-based analysis by definition. The benefits are obtained by shortening gene extensions where recombination is high and by lengthening them where recombination is low.

These gene definitions are not intended for interpretation of top association results in individual GWAS. In every case that a more detailed analysis is worth the extra effort, it should be done. Our choice of two GWAS as examples for illustration of the differences between the two gene definition approaches was motivated by the quality and reproducibility of GWAS, not to predicate the use of gene definitions in this field. The two GWAS were selected because they were of high quality, have found a large number of loci, have done detailed analysis of all the loci and have provided a full description of the genes selected for each of them. These are uncommon characteristics and we were fortunate that the two studies used different approaches for selecting the putative functional genes allowing a more thorough comparison of the  $\pm 50$  Kb and the  $SRR \geq 2$  gene definitions.

Other gene extensions based on genetic distances are possible. We chose the threshold of  $SRR \geq 2$  because it produced extensions of similar median physical length as the most used in previous studies. It will be inappropriate

and arbitrary to compare other SRR thresholds with the  $\pm 50$  Kb gene definition because these definitions will have different coverage of the genome and such comparisons will mix two components: differences in coverage and lack of correspondence between genetic and physical distances. By using the  $SRR \geq 2$  rule we assured an equivalent coverage of the genome and the comparison was focused in the lack of correspondence between the two distances. Later we found that it led to concordant results with detailed *post-hoc* analysis in 17 of 18 WTCCC GWAS associated loci and to inclusion of 90% of 43 functional candidates for red blood cell associated loci. Therefore, this definition seems convenient although we do not claim that more appropriate SRR thresholds could not be found for specific applications at around this value.

Our approach of using genetic distances in place of physical ones is widely applicable; but the gene definitions we provide are only directly applicable to Europeans. Other maps and specific genetic parameters will be necessary to study other ethnic groups. A genetic map for individuals of African ancestry has already been reported [22]. In addition, we have taken genetic distances from the deCODE recombination map [16], but genetic maps based on the decrease of LD can be taken as alternatives. Currently the best of these maps has been produced with HapMap samples [17,18]. Although the recombination map and the LD based maps have a high degree of correlation, there are differences between them and some gene definitions will be discordant. Both maps were obtained on the NCBI36 genome assembly that has been replaced by more recent ones. However, conversion of the maps to current assemblies will decrease their accuracy and we consider that is more accurate to convert SNP data to the NCBI36 assembly (with liftover in the UCSC browser at <http://genome.ucsc.edu/cgi-bin/hgLiftOver>, Remap in the NCBI site at <http://www.ncbi.nlm.nih.gov/genome/tools/remap>, or a similar tool), perform definition of gene units with the  $SRR \geq 2$  rule and run the intended analyses with the gene- or pathway-units.

We used the RefSeq catalogue of protein coding genes for our analysis [23]. At least other four human gene sets are widely available, all of them different in some aspect although sharing sources of information and methodologies and being more or less interconnected [24,25]. These sets are in continuous revision to incorporate findings of new experiments and technologies and none claims to be complete or definitive. The RefSeq set has been manually curated after incorporating information from multiple sources. It is considered conservative and trusted and other annotation projects use it as one of their inputs. Among those using RefSeq input, the GENCODE set combines manual and automatic annotation

and is more comprehensive by including the transcripts detected in the ENCODE project. However, the number of the RefSeq, UCSC and GENCODE protein coding genes is very similar. Differences between these sets are remarkable only in the number of transcripts per gene and in the number of exons for each gene [24]. For example, the number of transcript per gene is much larger in GENCODE than in RefSeq, with the UCSC set in between. These differences could slightly modify the boundaries of the gene units defined taking the RefSeq set as reference. Therefore, we consider that the provided gene definitions are generally valid and will perform well but would be not fully consistent with other gene sets.

## Conclusions

A definition of genes based on the coding sequence plus extensions whose length is given by genetic distances was shown to lead to more accurate results in the two sets of top association signals analysed. Use of this definition is made as simple as the commonly used until now by the list of gene coordinates on the physical map that is provided.

## Methods

### Baseline gene definitions

The RefSeq collection (UCSC RefSeq hg18) of 18 022 human protein-coding genes in autosome chromosomes and their map positions from the NCBI36 genome assembly, which corresponds to the used by the deCODE recombination map [16], were taken as the bases to which extensions were attached.

### Recombination information

Relative frequency of recombination for each 10 Kb bin of the human genome was obtained as standardized recombination rates (SRR) from the sex-averaged deCODE recombination map [16]. SRR are the result of dividing the recombination rate corresponding to each bin by the overall recombination mean for the genome.

### Gene definitions

A SRR given gene extensions of median length approaching the most commonly used 50 Kb boundary was searched. SRR inside the coding sequence were not considered. Two gene definitions were compared: one based on physical distance and the other based on genetic distance. The first included RefSeq sequences + 50 Kb in each direction; the second, the RefSeq definition extended in 10 Kb bins until the cumulative SRR that gave a similar median length. In this way, two extensions were generated for each RefSeq gene per definition, one for each end, 5' and 3'. Genes placed near the telomeres and the centromere of each chromosome were incompletely covered in the

deCODE recombination map. For the 2 669 genes in this situation, length of extensions based on genetic distances was made equal to the median length of all other extensions for this specific chromosome. These manipulations were done with PERL and Unix scripts that combined data from RefSeq and the deCODE recombination map, established the extension limits and generated the tabulated plain text file including one row per gene, and the columns: chromosome, "left" boundary, "right" boundary and gene name, which is available for download as supplementary material in Additional file 2 [20].

### Assessment of the gene definitions

Differences in length of the extensions between the  $\pm 50$  Kb and the SRR based definitions were calculated for all genes. In addition, the two gene definitions were applied to two large GWAS identifying multiple loci and using different criteria to highlight associated genes. Firstly, we examined 18 loci associated in the WTCCC GWAS [19]. This study examined about 14 000 patients, 2 000 of each of 7 different major diseases, and 3 000 healthy controls. Findings included 18 independent associations with  $p < 5 \times 10^{-7}$  analysed with detail in Figure five of the Nature paper. The authors defined regions of association (indicated by dotted vertical lines) extending until p values of the SNPs returned to background levels and, where possible, to recombination hotspots. We took these 18 loci and generated lists of genes overlapping with the association regions considering that any part of the genes according to the  $\pm 50$  Kb or the SRR definitions falling in between the dotted vertical lines was sufficient to count it. Secondly, we considered the 43 functional candidate genes in the 75 loci identified in the GWAS from van der Harst *et al.* analyzing 6 red blood cell quantitative phenotypes in 135 000 subjects [21]. These functional candidates were selected with two criteria: presence of nsSNP or of eQTL with  $r^2 > 0.8$  with the top associated SNP. One of the loci was excluded from our analyses due to its exceptionally high LD and unique inheritance pattern due to a common inversion polymorphism under positive selection in Europeans of about 900 Kb in chromosome 17q21 [26]. We took the remaining 43 functional candidate genes and tested if they would be included among the highlighted genes using the  $\pm 50$  Kb or the SRR gene definitions around the top associated SNP. The study of van der Harst *et al.* included other criteria for prioritizing candidate genes based in analysis of the bibliography or in the physical proximity that we did not consider here. Comparisons of the number of genes failing with each of the two gene definitions were done with the one-tailed Fisher exact test applied to the 2x2 contingency tables.

## Additional files

**Additional file 1: Table S1.** List of genes included in each of the top associated loci from the WTCCC GWAS (Ref 19) by the authors of the study, with the  $SRR \geq 2$  definition and with the  $\pm 50$  Kb rule. This file is available in the Dryad Digital Repository, doi:10.5061/dryad.p58hb, <http://doi.org/10.5061/dryad.p58hb>.

**Additional file 2: Coordinates with the boundaries for the  $SRR \geq 2$  gene definitions.** Tabulated plain text file including one row per gene, and the columns: chromosome, "left" boundary, "right" boundary and gene name. This file is available in the Dryad Digital Repository, doi:10.5061/dryad.p58hb, <http://doi.org/10.5061/dryad.p58hb>.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

CR-F, MC and AG contributed to the conception and design of the study, analysis and interpretation of data, drafting the article and final approval of the submitted version. AG takes responsibility for the integrity of the work as a whole, from inception to finished article. All authors read and approved the final manuscript.

## Acknowledgements

The present work was supported by Fondo de Investigacion Sanitaria of the Instituto de Salud Carlos III (ISCIII) of Spain, grant PI11/01048 that is partially financed by the European Regional Development Fund of the European Union. CR-F is a pre-doctoral fellow of the ISCIII and MC has been supported by funds of the Xunta de Galicia (Isabel Barreto Program). The funding bodies have no role in design, analysis, and interpretation of data or in the writing of the manuscript and in the decision to submit the manuscript for publication.

Received: 22 October 2013 Accepted: 20 May 2014

Published: 28 May 2014

## References

- Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S: **A versatile gene-based test for genome-wide association studies.** *Am J Hum Genet* 2010, **87**:139–145.
- Nyholt DR, Low SK, Anderson CA, Painter JN, Uno S, Morris AP, MacGregor S, Gordon SD, Henders AK, Martin NG, Attia J, Holliday EG, McEvoy M, Scott RJ, Kennedy SH, Treloar SA, Missmer SA, Adachi S, Tanaka K, Nakamura Y, Zondervan KT, Zembutsu H, Montgomery GW: **Genome-wide association meta-analysis identifies new endometriosis risk loci.** *Nat Genet* 2012, **44**:1355–1359.
- Tang W, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD, Samani NJ, Basu S, Gogele M, Davies G, Lowe GD, Tregouet DA, Tan A, Pankow JS, Tenesa A, Levy D, Volpato CB, Rumley A, Gow AJ, Minelli C, Yarnell JW, Porteous DJ, Starr JM, Gallacher J, Boerwinkle E, Visscher PM, Pramstaller PP, Cushman M, Emilsson V, Plump AS, et al: **Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease.** *Am J Hum Genet* 2012, **91**:152–162.
- Vijai J, Kirchhoff T, Schrader KA, Brown J, Dutra-Clarke AV, Manschreck C, Hansen N, Rau-Murthy R, Sarrel K, Przybylo J, Shah S, Cheguri S, Stadler Z, Zhang L, Paltiel O, Ben-Yehuda D, Viale A, Portlock C, Straus D, Lipkin SM, Lacher M, Robson M, Klein RJ, Zelenetz A, Offit K: **Susceptibility Loci associated with specific and shared subtypes of lymphoid malignancies.** *PLoS Genet* 2013, **9**:e1003220.
- Zhang K, Cui S, Chang S, Zhang L, Wang J: **i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study.** *Nucleic Acids Res* 2010, **38**:W90–W95.
- Fridley BL, Biernacka JM: **Gene set analysis of SNP data: benefits, challenges, and future directions.** *Eur J Hum Genet* 2011, **19**:837–843.
- Wang K, Li M, Bucan M: **Pathway-based approaches for analysis of genomewide association studies.** *Am J Hum Genet* 2007, **81**:1278–1283.
- Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L: **Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data.** *Am J Hum Genet* 2010, **86**:860–871.
- Wei S, Wang LE, McHugh MK, Han Y, Xiong M, Amos CI, Spitz MR, Wei QW: **Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility.** *Carcinogenesis* 2012, **33**:1531–1537.
- Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A: **Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations.** *PLoS Genet* 2012, **8**:e1002714.
- Liu Y, Xu H, Chen S, Chen X, Zhang Z, Zhu Z, Qin X, Hu L, Zhu J, Zhao GP, Kong X: **Genome-wide interaction-based association analysis identified multiple new susceptibility Loci for common diseases.** *PLoS Genet* 2011, **7**:e1001338.
- Siontis KC, Patsopoulos NA, Ioannidis JP: **Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies.** *Eur J Hum Genet* 2010, **18**:832–837.
- Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N: **Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder.** *Am J Hum Genet* 2009, **85**:13–24.
- Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJ: **Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways.** *Hum Mol Genet* 2011, **20**:3494–3506.
- Matisse TC, Chen F, Chen W, De La Vega FM, Hansen M, He C, Hyland FC, Kennedy GC, Kong X, Murray SS, Ziegler JS, Stewart WC, Buyske S: **A second-generation combined linkage physical map of the human genome.** *Genome Res* 2007, **17**:1783–1786.
- Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Gylfason A, Kristinnsson KT, Gudjonsson SA, Frigge ML, Helgason A, Thorsteinsdottir U, Stefansson K: **Fine-scale recombination rate differences between sexes, populations and individuals.** *Nature* 2010, **467**:1099–1103.
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, et al: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851–861.
- Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome.** *Science* 2005, **310**:321–324.
- Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**:661–678.
- Supplementary data from: *Genetic distance as an alternative to physical distance for definition of gene units in association studies.* [<http://doi.org/10.5061/dryad.p58hb>]
- van der Harst P, Zhang W, Mateo Leach I, Rendon A, Verweij N, Sehmi J, Paul DS, Elling U, Allayee H, Li X, Radhakrishnan A, Tan ST, Voss K, Weichenberger CX, Albers CA, Al-Hussani A, Asselbergs FW, Ciullo M, Danjou F, Dina C, Esko T, Evans DM, Franke L, Gogele M, Hartiala J, Hersch H, Holm H, Hottenga JJ, Kanoni S, Kleber ME, et al: **Seventy-five genetic loci influencing the human red blood cell.** *Nature* 2012, **492**:369–375.
- Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ, Bock CH, Boerwinkle E, Cai Q, Caporaso N, Casey G, Cupples LA, Deming SL, Diver WR, Divers J, Fornage M, Gillanders EM, Glessner J, Harris CC, et al: **The landscape of recombination in African Americans.** *Nature* 2011, **476**:170–175.
- Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucleic Acids Res* 2012, **40**:D130–D135.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, et al: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22**:1760–1774.
- Harte RA, Farrell CM, Loveland JE, Suner MM, Wilming L, Aken B, Barrell D, Frankish A, Wallin C, Searle S, Diekhans M, Harrow J, Pruitt KD: **Tracking and**



coordinating an international curation effort for the CCDS project.

*Database (Oxford)* 2012, **2012**: doi:10.1093/database/bas008.

26. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K: **A common inversion under selection in Europeans.** *Nat Genet* 2005, **37**:129–137.

doi:10.1186/1471-2164-15-408

**Cite this article as:** Rodriguez-Fontenla *et al.*: Genetic distance as an alternative to physical distance for definition of gene units in association studies. *BMC Genomics* 2014 **15**:408.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

